Neural Computing and Applications (2024) 36:18919–18934
https://doi.org/10.1007/s00521-024-10207-0

ORIGINAL ARTICLE

KDTL: knowledge-distilled transfer learning framework for diagnosing mental disorders using EEG spectrograms

Shreyash Singh[1] · Harshit Jadli[1] · R. Padma Priya[1] · V. B. Surya Prasath[2,3,4,5]

Received: 3 August 2023 / Accepted: 12 July 2024 / Published online: 1 August 2024

Abstract

Electroencephalography (EEG) is a well-known modality in neuroscience and is widely used in identifying and classifying neurological disorders. This paper investigates how EEG data can be used along with knowledge distillation-based deep learning models to detect mental disorders like epilepsy and sleep disorders. The EEG signals are converted into time–frequency plots using short-time Fourier transforms. Further, we propose a novel methodology for using knowledge distillation-based transfer learning (KDTL). Knowledge distillation is becoming quite prevalent in the machine learning field and is associated with various applications in our work; we propose its use in the detection of mental disorders from EEG spectrograms. We convert the EEGs using short-term Fourier transform to obtain time–frequency representation and apply teacher-student by first training a large teacher model and use knowledge distillation to train a student model. In our experiments, we found that ConvNext teacher and MobileNet student combination obtained better results. Our proposed KDTL approach is tested on two datasets with multiple cases, namely the Bonn and ISRUC datasets and obtain 98% and 95% accuracies, respectively. Further experimental results show that the overall KDTL methodology can obtain high classification accuracy across both datasets in binary and multiclass classifications and proves to be better than multiple prior works. Further, our KDTL approach provides a way to train lightweight models which have a smaller number of trainable parameters and thereby constitute lower training time overall. Our proposed KDTL-based approach obtained accurate results in diagnosing mental disorders from EEG spectrograms. Compared to other related methods, KDTL outperformed across tasks with obtained good results in both binary and multiclass classifications.

Keywords EEG · Transfer learning · Knowledge distillation · Mental health · Spectrograms

1 Introduction

In today's world, mental disorders are quite prevalent; globally, an estimated 5 million people are diagnosed with epilepsy each year and hence there must be techniques available to detect these disorders. Epilepsy which is one of them is a chronic noncommunicable mental disorder which is associated with convulsions in the body. Some of the causes for it include brain damage from perinatal causes, stroke, infection, and tumors in the brain. It can cause

✉ R. Padma Priya
padmapriya.r@vit.ac.in

Shreyash Singh
write.shrey@gmail.com

Harshit Jadli
harshit.jadli2019@vitstudent.ac.in

V. B. Surya Prasath
prasatsa@uc.edu

[1] Department of Software Systems, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

[2] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA

[3] Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH 45257, USA

[4] Department of Biomedical Informatics, College of Medicine, University of Cincinnati, Cincinnati, OH 45267, USA

[5] Department of Computer Science, University of Cincinnati, Cincinnati, OH 45221, USA

severe issues since it can lead to intellectual problems. Depression with epilepsy is also very commonly observed, both in children and adults. As a result, the impacts can also range from mild to severe in various individuals and multiple causes have been proposed for it like, injury to the brain, irregular hormone levels, and usage of anti-seizure medications. According to the global campaign against epilepsy (GCAE), most cases of epilepsy start from childhood. It has also been observed that three quarters of patients in low-level income companies who suffer from epilepsy cannot afford the treatment available. Patients of epilepsy also have to deal with other challenges of unemployment, poverty, and getting a stigma from society which leads to them getting isolated by society. Another such disorder is sleep disorder. It can adversely affect a person's life and health. There are several types of sleep disorders like insomnia (difficulty in falling asleep), sleep apnea (abnormal stopping of breathing during sleep), narcolepsy, etc. It has been estimated that about 70 million people in the US are affected by some type of sleep disorder which makes it a vital field to study. It also affects the cognitive function of the patients, affects memory and can also cause depression. The severity of sleep apnea is classified using the apnea–hypopnea index (AHI). This usually ranges from normal to severe.

There are multiple methods which help in solving such issues as fMRI (functional magnetic resonance imaging), EOG (electrooculogram), PSG (polysomnography), and many more. Methods like PSG are complex, and research was conducted using ECG to detect sleep disorders. Another well-known modality is electroencephalography (EEG) with which we can gain acute insights into how to cure these disorders [1] and understand the causes of the same. Using EEG, we can look at the subject's mental activity [2], illness [3], and stress [4], see [5] for a recent review. EEG enables researchers to observe brain waves and detect abnormalities that could signal neurological conditions. It provides valuable insights into mental activity and potential remedies for sleep disorders. The conventional approach to finding a remedy involves inspecting 24-h EEG activity by trained neurologists which is very time-consuming and may involve errors also as such it is the most optimal approach available. Hence, it is necessary that research should be conducted in this field to address issues with robustness, and accuracy, and at the same time such solutions should be feasible so that they can be applied to real-time situations as well. By automating or enhancing EEG analysis, the field can develop feasible solutions suitable for real-time diagnosis, ultimately leading to quicker, more reliable, and cost-effective remedies. Continuing research in this area is crucial for advancing diagnostic techniques and improving patient outcomes.

A successful methodology involves providing a solution which can provide accurate solutions using all the features extracted from the EEG signals of the subject. For this, the researchers should initially focus on feature extraction and using these features to provide an efficient solution. In this study, the focus is on proposing a model, which is lightweight and at the same time is efficient enough to provide accurate results. The various methodologies used by scholars to extract features are presented in the next section. For this study, we have used the time–frequency features using spectrograms other methods involve using empirical wavelet transform [6] to detect epilepsy or using empirical mode decomposition (EMD) and discrete wavelet transform (DWT) [7] to get Shannon entropy, log-energy entropy, and Renyi entropy for characterizing the EEG signals. Recently, multiple attempts have been made to work with EEG data using deep learning models like convolutional neural network (CNN), sparse autoencoder (SAE), and machine learning methods like support vector machine (SVM) and random forests. George et al. [8] studied the classification and recognition of emotions, valence, and arousal in this case, based on EEG signals. The method uses a time–frequency feature for the classification. The optimal features are isolated and fed to the support vector machine (SVM). The dataset used was the DEAP dataset and the accuracy obtained was 92.36%. Zheng et al. [9] proposed the use of the Hilbert–Huang transform which led to the removal of usage frequency bands which are fixed to prevent the low accuracy in classifiers using EEG data. It removed the limitation of constant frequency and amplitude. The data used for the research was obtained using BrainVision wireless ActiCap 64-channel scalp EEG recording system and different classifiers were used like SVM, random forest, and XGboost and the accuracy was obtained to be between 77 and 96% for different feature sets. Tabar and Ugur [10] used a convolutional neural network and stacked autoencoder in order to classify the EEG motor imagery signals. The EEG data are used of C3, Cz, and C4 electrodes and the short-time Fourier transform is obtained of the 2 min duration signals with a window size of 64 and the average accuracy was obtained to be 74.8% in CNN and 57.7 in SVM. The reason for this low accuracy in the stacked autoencoder is caused by loss of neighborhood information in SAE.

Madhavan et al. [11] utilized the time–frequency matrices obtained using Fourier SST and wavelet SST to classify the focal and non-focal EEG signals. By using two-dimensional CNNs, the accuracy was found to be 99% and the comparison of the STFT (short-time Fourier transform) and discrete complex wavelet transform was done which revealed that STFT is better. Tawhid et al. [12] used EEG data-based autism classification using time–frequency

spectrogram images. The data are preprocessed using filtering and normalization, then the short-time Fourier transform is performed to get the time–frequency plots and then the principal component analysis (PCA) is performed to obtain the textural features and the data are fed to the support vector machine. The accuracy obtained here was 95.25%.

Chao et al. [13] used the concept of knowledge distillation to distill the knowledge from the multichannel teacher model to the single-channel student model and the accuracy observed in this case was 86.5%. Using different techniques of knowledge transfer (Domain Same-Channel (SDSC), Same-Domain Cross-Channel (SDCC), Cross-Domain Same-Channel (CDSC), and Cross-Domain Cross-Channel (CDCC)) and Seqsleepnet model as a teacher and student model KD is implemented. Khan et al. [14] utilized the concept of knowledge distillation [15] to classify ADHD patients from healthy patients. The dataset used is ADHD-200. Initially, the larger model, resNet64, is trained using the feature vectors extracted and after this, the knowledge distillation is performed, and finally sequential forward feature selection is performed to obtain the most discriminating features. It was observed that using this technique an accuracy of 60% to 81% is obtained. Knowledge distillation (KD) and EEG classification's application has been demonstrated in the field of affective computing. The training of the student model is done with the help of the data from the teacher model with the help of capsule networks which further promotes the reduction in size of the model without an increase in loss as demonstrated in [16] in which the deep learning model used is LSTM and two different EEG datasets are used (SEED, SEED-VIG). Ieracitano et al. [17] utilized wavelet transforms to obtain the time–frequency plots to obtain features of different frequency bands and used different classifiers (LR, MLP, SVM) to perform the classification of the Alzheimer's disease, along with this they have also made use of bispectrum analysis to observe the state of the brain of different patients. Using both types of analysis the authors demonstrated that by using wavelets and bispectrum analysis, we can use the information of high-order spectral features more effectively. In our work, here we propose a compact model using knowledge distillation which can be run on edge devices present in medical facilities due to its low size.

Convolutional neural networks (ConvNets) are very popular in deep learning due to their usage of various types of filters that can capture spatial properties in input data [18]. This particular class of deep learning models is quite often used in the applications of object, class, and category recognition in imaging data. Neural networks consist of hidden layers composed of neutrons which are connected to the neutrons of the previous layer and pass the value based on the activation functions and through the last layer we obtain the class score. ConvNet consists of a similar methodology with the difference being that here the inputs used are images since with a normal neural network there exists the problem of having a large number of parameters even for a small-resolution image. Here, the neurons have three dimensions: height, weight, and depth. The filters are slid over the width and height of the input and the dot products are computed, a 2-dimensional activation map is generated which contains the outputs of the filter and eventually the architecture learns the filters to activate upon seeing a visual feature. However, there is an issue related to the ConvNet, as more and more layers are added to our network the training and testing error rates also increase, called the problem of vanishing gradient, and for this reason, residual networks (ResNets) were introduced in 2015 [19]. ResNets solves this issue using the concept of skip connections. The activation of a layer is connected to a subsequent layer by skipping a few of the intermediate layers forming a residual block. The stack of these blocks is called ResNets. Liu et al. [20] explored the improvements in ResNets and proposed the ConvNext model. Tao et al. [21] have demonstrated cotton disease detection using the ConvNext model along with a multiscale spatial pyramid attention (MSPA) module on two different datasets. The accuracy obtained was 97.2–99.7%. Fan et al. [22] using the ConvNext model performed low light image enhancement by proposing a novel attention ConvNext Module by combining two ConvNext modules and merging their features. ConvNets have been incorporated into other neural networks [23] including the region-based mask R-CNN [24]. We summarized the prior works in applying machine learning and deep learning models for EEG spectrograms in Table 1. Based on this literature review and the limitations of these prior approaches, we have decided to use MobileNet for the student model, since it has a lower number of parameters and less computational cost required and hence can potentially be used in mobile devices. We have tested our KDTL methodology on two publicly available datasets and aim to provide insights which can be used by researchers in this field to gain insights and inferences.

The remaining paper is divided into the following sections. Section 2 introduces the KDTL methodology on EEGs. Section 3 provides detailed experimental results of the study and Sect. 4 concludes the paper with future prospects.

Table 1 Related works from the literature and their corresponding advantages and limitations

Paper	Proposed method	Advantages	Limitation
[8]	The different frequency bands are used to differentiate between arousal and valence emotional states using SVM	Using SVM after feature selection using box and whisker plots helps in reduce the dimensionality of the input space and also provides high accuracy (93.26%)	The work considered only statistical features and classified only 2 emotions: valence and arousal
[9]	The time–frequency plots are obtained using Hilbert–Huang transform to improve the accuracy of the classifier	The limitation of having a constant frequency and amplitude is removed which is present in the Hilbert transform normally used	The accuracy is 70–90% which is not a lot as there have been methods found which give higher accuracy than conventional methods
[10]	In this paper, CNN and SAEs are used to perform differentiation in EEG motor imagery signals	Using CNN, the features of the time, frequency location information are successfully obtained	Low accuracy due to loss of neighborhood information
[25]	In this paper, the concept of knowledge distillation is used to demonstrate how the size of the dataset can be effectively reduced and how using a simpler model we can achieve good results in real time	Knowledge distillation allows successful training of a simpler model from a complex model which cannot be used in real-life scenarios	The generalization proposed is not verified as the number of classes is a lot and the number of images is also in huge number. The self-learning KD also tends to use a lot of computational resources
[11]	Classification of focal and non-focal EEG signals using SST and CNN	Better than DT-CWT, STFT, and other conventional time–frequency generation techniques, with the highest accuracy being 99.94%	The CNN involves a lot of computational resources unless GPU is utilized making it difficult to use on mobile devices
[13]	Knowledge distillation is used to train a single-channel student model from a multichannel teacher model in order to classify sleep stages	Using knowledge distillation, an improvement in accuracy of 2% was observed in the student model	The accuracy is lower as compared to the other proposed methods
[26]	Classification of ASD and epilepsy is performed. DWT and Shannon entropy and the largest Lyapunov exponent are used to extract features	The highest accuracy of 94.6% using DWT, Shannon entropy, and KNN	The number of samples is low which limits the generalizability of the findings to a larger population
[27]	Image fusion model using knowledge distillation and explainable AI module-based generative adversarial network with dual discriminators	Using knowledge distillation, the size of the overall dataset is reduced from 218.34 to 66.03 MB and high accuracy of 97.63% was observed	In 64 × 64 and 128 × 128 overfitting was occurring
[17]	Bispectrum analysis and time frequency plots are used to obtain the representation of the EEG signals and are tested with MLP, SVM, AE, and LR	MLP was found to consume less computational resources than other conventional approaches which had a very high accuracy of 96%	A large sized dataset was not tested which could lead to some features being missed for classification
[22]	Low light image enhancement is performed by merging 2 convnext modules using Selective Kernel Attention Module (SKAM)	Compared to conventional methods like RetinexNet, MBLLEN, and KinD much better results are obtained showing much better preservation of details	High-noise images do not give satisfactory results, and a loss of details is observed
[28]	MobileNet is used to detect the disease and is compared with state-of-the-art approaches like ResNet152 and InceptionV3	The accuracy of the model was almost the same as the conventional approaches. The efficiency of the MobileNet was also high due to depthwise separable convolution and parameter sharing	The accuracy was low (72%), due to a low number of images (2004)
[29]	Dense MobileNets are proposed with convolutional layers of the same size as the input feature map as dense blocks to further decrease the computational resources and parameters	The classification accuracy is improved by the proposed method to 92%	The improvement of accuracy was only about 1.3% and, in a few cases, it was observed to have a lower accuracy although just 0.4% as compared to the standard MobileNet model

2 Proposed knowledge-distilled transfer learning (KDTL) framework

2.1 KDTL framework for EEG spectrograms

In this work, we proposed a novel hybrid DL framework called knowledge-distilled transfer learning (KDTL) to perform classification problems with the help of time frequency-based images, see Fig. 1. The idea here is to use spectrograms to classify various mental disorders, which aims to give a contribution to the field of EEG and deep learning. We have performed various classifications using two different datasets to judge how well our model works. There are three stages in total: (1) the EEG data are converted to spectrograms using short-term Fourier transform. In doing so, we can obtain the time–frequency. (2) Once the spectrograms are obtained, we are building and training our teacher model (ConvNext). (3) After the teacher model is trained, we use knowledge distillation to train the student model (MobileNet) using the soft predictions of the teacher and the hard predictions. We next provide the details of our proposed KDTL framework and the EEG datasets utilized.

2.2 Datasets used

The first dataset used is the EEG public dataset from Bonn University containing normal and epileptic patients. There are 5 sets present each consisting of 100 signals of 23.6 s duration. The sampling rate given is 173.61 Hz. The A and B set present consists of EEG recordings from a healthy patient with eyes open in set A and eyes closed in set B. The sets C, D, and E contain the data from epileptic patients. Table 2 provides the breakup of the dataset, and Fig. 2 shows example signals from the different subsets.

The second dataset used is the ISRUC dataset which contains the EEG data present in 3 groups which contains 100 adults with disorder, 8 with disorder, and 10 healthy adults of age between 30 and 58 years. The data is extracted from 6 EEG channels C3-A2, C4-A1, F3-A2, F4-A1, O1-A2, and O2-A1 sampled at 200 Hz. 2 experts evaluate the recordings using AASM rules and have classified the sleep stages also. On these signals we perform STFT in order to obtain the time frequency plots of the signals which are used for training our teacher model, for example, ConvNext.

For our study, we have used 80% training data, 10% validation data, 10% testing data. There are 4096 samples in each set and the total number of spectrograms generated is 500, and the total number of samples in ISRUC are more than 15,000. In the Bonn dataset, 300 images were used for training, 100 for testing, and 100 validation, and in ISRUC 13740, spectrogram images were used for training, 1716 for validation, and 1728 for testing. To ensure consistency in our results, we preprocessed all images to a uniform size of 224 × 224 before initiating the training process.

2.3 Time–frequency representation

Time–frequency representation allows us to perform the analysis of the features of the nonstationary signals in the

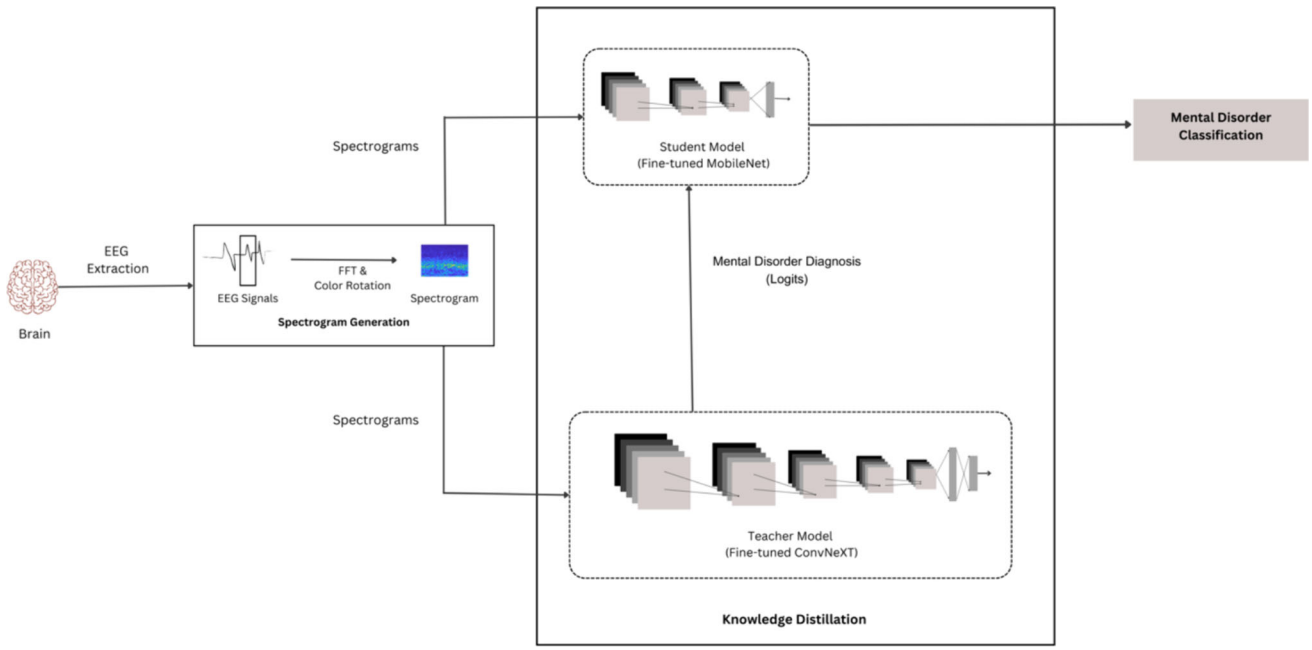

Fig. 1 Knowledge-distilled transfer learning (KDTL) framework considered in this work for analyzing the EEG-based spectrograms for diagnosing mental disorders

Table 2 Details of the Bonn University used in our experiments

Set	Subject	State of the subject
Set A	5 healthy volunteers	Awake and eyes open
Set B	5 healthy volunteers	Awake and eyes closed
Set C	5 patients	Seizure free
Set D	5 patients	Seizure free
Set E	5 patients	Seizure

time and frequency domain. There are different ways to compute the time–frequency representation. For example, using a complex Morlet wavelet, we use a Gaussian window and compute the complex Morlet wavelet by performing the dot product between a sine wave and the Gaussian window which is used as a kernel. The Fourier transforms of the input signal and the kernel are computed and multiplied since they are in the frequency domain and the inverse Fourier transform is taken to obtain the result of the convolution of the kernel and the input signal and these results can be plotted to obtain the time–frequency representation. Here, we have used short-time Fourier transform (STFT) to obtain the time–frequency plots of the EEG signals. From the original signal we cut out a small epoch present in the window, this smaller representation is tapered to attenuate the edges which are present in the beginning and the end and a Fourier transform is taken of this representation which gives us the power spectrum of this signal, this representation is "rotated" giving a single column in time frequency plots having a frequency in the x-axis and time in y-axis and power in the z-axis. The window then slides across the signal and the process is repeated to obtain the complete time–frequency plot one-time point at a time overall frequency. There is a trade-off present between time and frequency as we change the window's length. To counter this trade the window can be changed for different ranges of frequency. We have used the Han window as our windowing function which is 200 points long and the sampling rate has been taken as 173 Hz for the first dataset. The formula for STFT is given by

$$F(w,t) = \int_{-\infty}^{\infty} f(t)\psi^*(t-\tau)e^{-\mathrm{jwt}}\mathrm{dt},$$

where $\psi(t)$ is the window function. In this manner, we have obtained the spectrograms from our data and we will be using these images of time–frequency representation to perform the classification.

2.4 Teacher model

In our work, we have used the ConvNext model to train our teacher model. One of the reasons for choosing ConvNext is inductive bias. Inductive bias subsumes the assumptions which are required to perform a prediction. In the case of the ConvNext model, there are multiple presents which make it overall much more favorable for the computer vision domain. It also facilitates the sharing of computations when used in the sliding window approach as shown in [30]. Depthwise convolution is used and along with

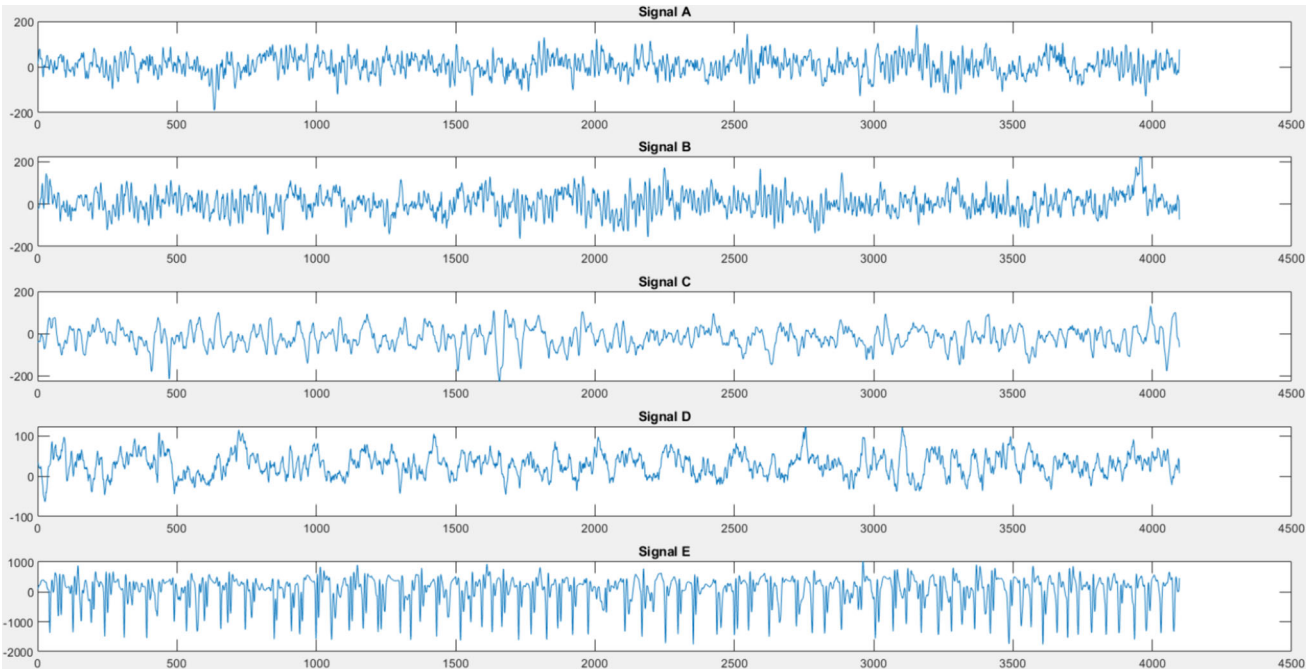

Fig. 2 Examples of signals from the Bonn University dataset

1×1 convolution, it helps in separation of spatial and channel mixing which further reduces the FLOPs while maintaining the accuracy. The kernel size is also increased since it provides an increase in accuracy along with reduced FLOPS. The kernel size is 7×7 in our case. The ReLU which is commonly used as the activation function is replaced with GeLU which helps in avoiding overfitting which is used in many transformers like Google BERT and the number of activation functions and normalization layers are also reduced. LN (Layer Normalization) is utilized to prevent the mini-batch distribution issue.

$$\text{GELU}\ (x) = \phi(x) * x$$

$$\text{LN}(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \int}} * \gamma + \beta$$

Here, x is the input feature map, ε is a constant so that the dominator does not converge to 0, μ, σ are the mean and standard deviation of the channel values, * is the scalar multiplication, γ, β denotes the learnable scale and offset factor and $\phi(x)$ is the cumulative distribution factor. We have prepared 5 different ConvNext models standard, large, small, tiny, and extra-large to obtain the different results. The last layer of the model consists of 2 classes and the rest of the layers have been obtained from the pre-trained model present in Keras. This last layer is trained using the data obtained above and the activation function used is the SoftMax function. In this manner only the weights of the final layer are modified and not the rest of the layers. Figure 3 shows the overall architecture considered here. The optimizer used is the Adam optimizer [31] which combines the advantage of gradient descent with momentum and the RMS prop algorithm. The learning rate is 0.0001 and the loss function used is sparse categorical cross-entropy. We have used early stopping which restores the configuration of the model to the configuration present during the best epoch. The teacher model is then trained for 10 epochs. The accuracy obtained from our teacher model is found to be 100%. In this study, while the observed outcomes may initially suggest overfitting, it is important to note that this may be due to a high number of epochs. This hyperparameter set was obtained after extensive experiments with hyperparameter tuning and optimization.

2.5 Student model and knowledge distillation

For the student model, we have used MobileNet. They are a class of lightweight deep convolutional neural networks which are smaller in size as compared to the other models. The MobileNet makes use of depthwise convolution along with 1×1 convolution. The depthwise convolution is applied to each of the channels, and 1×1 convolution is then performed to combine the output of the depthwise convolution. This factorization of using separate layers for filtering and combination reduces the computation and size by a huge amount. The cost of computation involved in MobileNet is given by,

$$S_k.\ S_k.\ I.\ S_f.\ S_f.$$

This computational cost is drastically less compared to standard convolution [32] and is given by,

$$S_k.\ S_k.\ I.\ O.\ S_f.\ S_f.$$

Here, I is the number of input channels, O is the number of output channels, $S_k \times S_k$ is the kernel size, and $S_f \times S_f$ is the feature map size. The depthwise convolution filters from input channels and for the combination of the features an additional layer is needed through 1×1 convolution to generate the features. The 3×3 depthwise convolution used gives us a reduction of 8–9 times of computation resources. Using MobileNet we can compose even smaller models if a requirement is needed. For that, we can modify the width multiplier value of alpha which thins the network at each layer. For our work, we have used the default value of 1 as present in the baseline MobileNet, however, if required, reduced MobileNet can also be implemented by reducing the value of alpha. There is a trade-off for

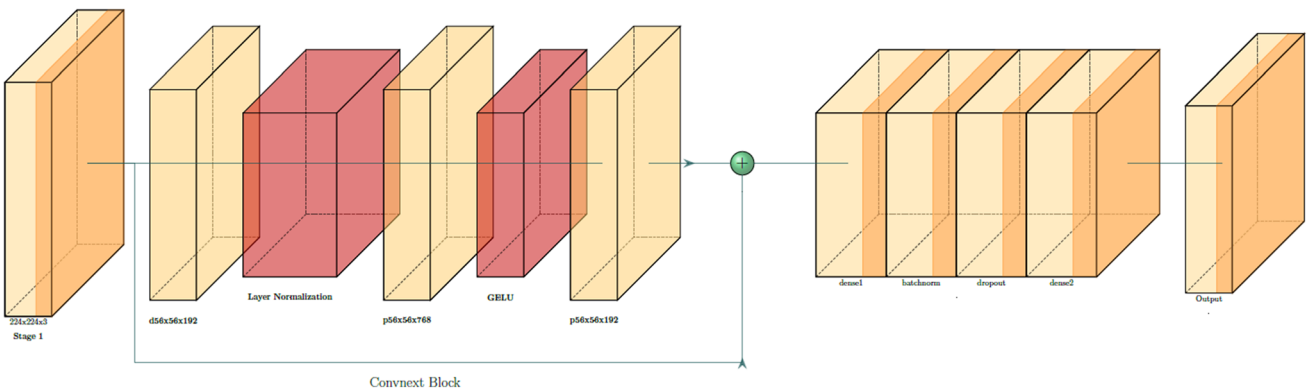

Fig. 3 The architecture of ConvNext in which d refers to depthwise convolution, where *p* refers to pointwise convolution

accuracy also involved however it is not by a huge amount. For example, Inception V3 had an accuracy of 84% and MobileNet with alpha as 1 had an accuracy of 83.3%, MobileNet also performs better than AlexNet with a 4% better accuracy for the Stanford Dogs dataset. Multiple works have been done using MobileNet, in [33] authors have used MobileNet and then used knowledge distillation to reduce performance loss and demonstrated a reduction of 30% complexity in computation as compared to the standard methods in automatic scenario recognition in vehicle-to-vehicle systems. Chiu et al. [34] combined MobileNet with a Feature pyramid network to propose a novel object detection method with 75% accuracy and the ability to function in a resource-constrained environment, i.e., in embedded platforms. In [28, 35], MobileNet has been used for the classification of apple leaf diseases and flower recognition with the accuracy of 73.5% and 88.37%. In [28], authors have also compared the results with InceptionV3 and ResNet152 both displaying accuracy of 75% and 77.65%. This lower accuracy can be explained because the MobileNet model, being lightweight, uses a lower number of parameters due to which it is unable to identify abstract features as compared to the other models. Wang et al. [29] proposed two MobileNet models by introducing dense blocks to increase the feature maps produced using less convolutional kernels and by reducing the growth rate have further reduced the computation costs by nearly half, the accuracy found was 96%.

There have been many attempts made to produce models to work in real time due to constraints due to latency and throughput. One of the solutions to these limitations is knowledge distillation. The idea behind it is that a complex "parent" model is used to capture the information and then "distill" the information in a smaller model which is much easier to deploy in real time. Due to this reason, it is becoming quite popular in the fields of image recognition, speech recognition, and NLP [13]. The information is used from the teacher model by reducing the difference between the logits of the complex model and those produced by the simple model. There is however an issue in this process, the softmax function output obtained from the teacher model has a very high value of the correct class as compared to the other classes, as a result, the information other than the fundamental truth which is already known to us is not obtained. For this purpose, the class probability is calculated using the formula:

$$p_i = \frac{e \exp\left(\frac{z_i}{\rho}\right)}{\sum_i \exp\left(\frac{z_i}{\rho}\right)}$$

where, z_i is the logits, and ρ is the temperature parameter. This "dark knowledge" is obtained from the teacher model and is used in the training of the student model. It helps in reducing the problem of overfitting in the student model. The loss of the student model is calculated using a comparison of the output of the student model and the fundamental truth values along with the teacher values.

$$L = \alpha * H(y, \sigma(zs)) + \beta * H(\sigma(zt; \rho))$$

Here, α is the width multiplier, H is the loss function, y is the ground truth label, and σ is the softmax function with the parameter of temperature ρ.

Knowledge distillation can be performed using multiple methods. It can be performed based on the number of teachers, based on data format. KD based on data format is essentially used when the unlabeled data for our simpler model is not present. The reason why we did not go for KD based on data format is that the results obtained are not realistic which is not suitable in various semantic segmentation as can be seen in [36]. The training and computation are also much more complicated. The complete rate of success in the case of using this method for many tasks like image super-resolution needs to be explored more. This is why we are performing KD based on teachers, with a single teacher in our case. In the case of a single teacher KD, there is a transfer of information from a complex teacher to a smaller model using feature data. In [26], the process of classification using EEG data is explained, the discrete wavelet transform is obtained, and power spectral density is plotted, and the different frequency bands are taken as the feature vectors and are fed to the SVM, KNN, and ANN models. While this process is very effective and displays a high accuracy of 97%, the entire process is too cumbersome and computationally strenuous. This is where knowledge distillation can be used because it would allow us to work with data in real time using a lightweight model as can be seen in [25] in which different learning methods (base training, standard knowledge distillation, reverse distillation, defective KD, and a self-training knowledge distillation) are evaluated, and it was observed that self-learning KD can be used in cases when there is difficulty in searching for an appropriate teacher. Another advantage of knowledge distillation is that it also allows us to effectively reduce the size of the dataset required by using a complex model and training the simple model with the same feature extraction power as the former. The results of the proposed method were evaluated using 5 techniques: spatial frequency, structural similarity, edge information transfer factor, normalized mutual information, and nonlinear correlation information entropy, and the conclusion obtained was that the proposed hypothesis is correct by having better metrics than the conventional methods.

However, there are a few limitations present to the knowledge distillation approach also, Malinin et al. [37]

have shown that knowledge distillation is not always the best approach since if the student model and the teacher model are too different from each other there is a degradation in the performance observed and hence it is very important that we choose appropriate model and only the necessary information should be transferred. In fact, although the proposed methodology for KD was that it could be used on any dataset it has been shown in [38] that there are a few datasets in which this is harder to achieve and it is important to choose the models carefully because sometimes the simpler model ends up compromising the KD loss at the cost of cross-entropy loss or vice versa which is why the authors proposed the early stopping of the student model. Further, transformation of hints and the transformation of the guided features is performed, hints are the values obtained from the teacher model which are used to guide the student model. Here, we alter the dimension of the feature representation of the teacher and make the student's feature match the dimension of the teacher's feature. Next, we determine the distance metric in order to evaluate our knowledge distillation. There are multiple methods used to evaluate this parameter, and typically they tend to use distance function, which is based on L1, L2 distance [39]. For example, in [40, 41] L2 distance has been used and in [42] L1 distance has been used. In our research, we have used KL divergence loss to measure the distance. Forcing the divergence between the self-attention probability distributions to be as small as possible preserves the behavior in the student. It would also allow us to compress the "behavior" of the teacher in the student in a better way as compared to the other methods [43].

For our student model, as shown in Fig. 4 we have used MobileNet since it is a lightweight model with a smaller number of parameters and computational cost as compared to the larger models with a negligible loss in accuracy as already stated above. A custom classifier is added to our student model to change the class number based on number of classes and like the teacher model only the final layer is altered and is trained on the data. This model is trained using the logits of the teacher model using knowledge distillation as explained above. For this, we have created a distiller class. For our distiller class, we have used the Adam optimizer, the alpha value taken is 0.1, and the temperature is taken as 10. The distiller class takes the logits of the teacher model and computes the distillation loss using the prediction of the student model and the teacher model. The loss is then computed using the formula given above, after this the gradients are computed, and the weights are updated. The distiller class also contains a test function to compute the prediction, calculate the loss and display the metrics. The distillation loss is computed using the KL divergence loss function and the student loss function is a categorical cross-entropy function. The number of epochs is set at 10, and the knowledge distillation approach is illustrated in Fig. 5.

3 Experimental results and discussion

3.1 Setup, parameters, configuration

In this section, we show our proposed KDTL framework along with various combinations of DL models and comparisons with other related approaches are detailed. Our model was trained on 2 different publicly available datasets: the Bonn University EEG dataset (dataset A) and the ISRUC sleep EEG dataset (dataset B). To validate these datasets, we have used different combinations of teacher and student models also. They are: (i) ConvNext and MobileNet (ii) ConvNext and shallow CNN (iii) ConvNext and fine-tuned CNN, and (iv) ConvNextTiny and MobileNet. The shallow CNN is a combination of max pooling layers, batch normalization layers, depthwise convolutional

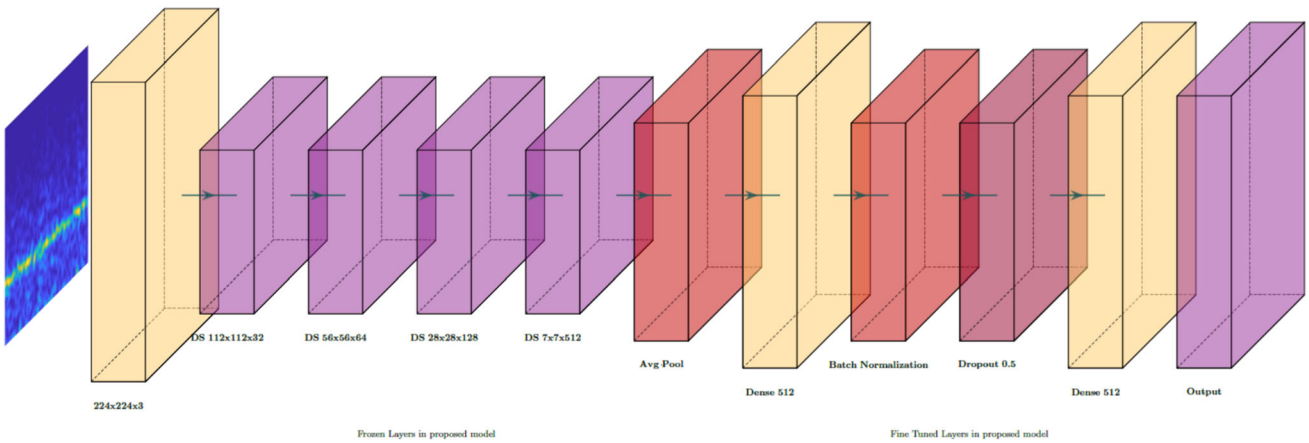

Fig. 4 Student model (MobileNet) architecture, in which DS refers depthwise separable convolution

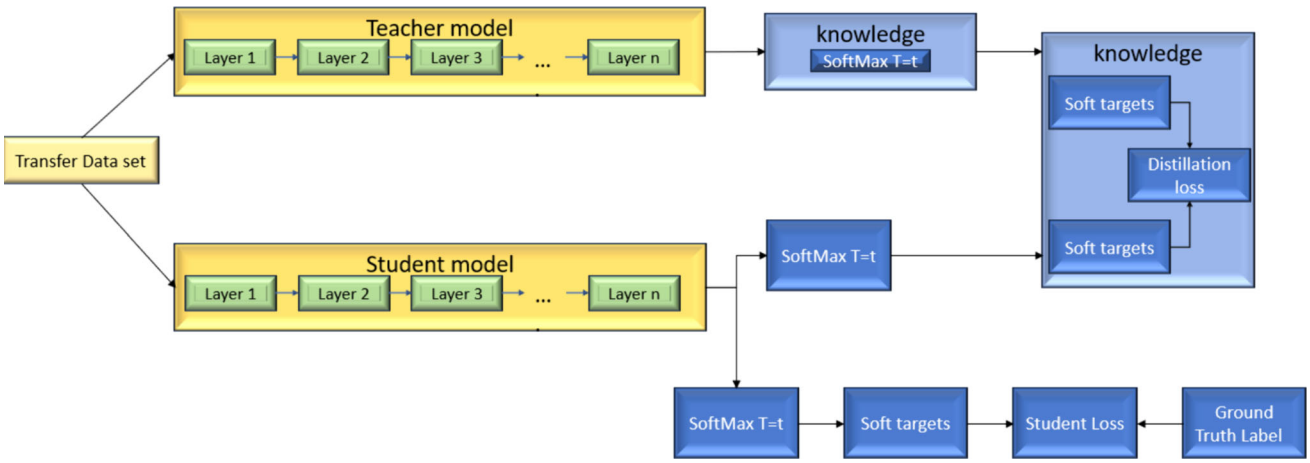

Fig. 5 Knowledge distillation approach utilized in our KDTL framework

layers, dropout layers, global average pooling layer, and dense layers and contains about 7000 parameters. The second CNN contains a max pooling layer, batch normalization, dropout layer, and dense layers and contains about 110,000,000 parameters. To assess the efficacy and performance of the proposed models, experimental evaluations were conducted on the widely used cloud-based platform, Google Colab with Nvidia K80 as GPU (12 GB memory) 0.82 GHz.

On the Bonn University dataset, we have classified 2 different cases: (1) 2 classes: healthy (Set A and B) versus epileptic (Set C, D, E) (2) 3 classes Set A, B versus Set C, D versus Set E. For the first case, we were able to obtain 100% accuracy and were able to outperform a lot of earlier works on the dataset and in the second case, the accuracy obtained was 98%. The various works done earlier had variable accuracy on multiple classes however in our work the accuracy was always within the range of 98–100%. Hence, our model is suitable for various classification applications, and the generalizing ability is also good. For the second dataset also the knowledge distillation helped in increasing the accuracy of the teacher model from 95 to 97% in classifying sleep disorder and reduced the number of parameters upon distillation which demonstrates the efficiency of our proposed methodology. The validation accuracy obtained was 97.28% which outperforms a lot of previous works done. For future works, we will investigate the performance of our proposed methodology on a greater number of classes on this dataset. Table 3 shows the proposed KDTL framework's parameter setup of various models. We obtained the optimal set of hyperparameters based on optimizing for higher accuracy.

3.2 Comparison of various DL models with the KDTL framework

Table 4 shows the different models we have implemented within the KDTL framework using the Google Colab framework with Python 3 Google Compute Engine backend (GPU), system RAM of 15 GB, and GPU RAM of 15 GB. The time taken is mentioned in the KDTL framework wherein we tested various DL models. We notice that the distilled time is typically < 1 s. We observed that the performance of shallow CNN was not satisfactory, which was expected since the model is rather simple and after fine-tuning the model the accuracy was higher and comparable to MobileNet which had a quarter of the number of parameters demonstrating the efficiency of the knowledge distillation process. Overall, our results demonstrate the effectiveness of knowledge distillation as a powerful tool for optimizing the performance of neural network models.

The binary classification also shows a high accuracy which beats the previous results. Upon increasing the number of classes to 5 the validation accuracy is 77%, this can be attributed to the fact that the extracted features do not take into account the regional connections between the different parts of the brain; there is room for improvement in the performance of traditional methods that use CNN and RNN as stated in [44]. As can be seen from Table 4, the training, testing, and validation accuracies demonstrate the trade-off between accuracy and model complexity. Based on the specific use cases, various combinations can be employed to achieve optimal results. We also notice similar performance in the multiclass classification using the ISRUC dataset, and our results show that Conv2Next

Table 3 Model parameters and configurations were tried within our KDTL framework

Model	Parameters	Configuration
ConvNext	196,230,336 (bi class) 196,456,131 (tri-class)	Added a dense layer
MobileNet	3,231,939 (bi class)	Added a dense layer
	3,231,939 (tri-class)	
Fine-tuned Conv2Next	234,769,090 (bi class)	Added a dense layer batch normalization, and a dropout layer
	234,769,603 (tri-class)	
Fine Tuned MobileNet	3,757,251 (bi class)	Added a reshape layer, dense layer batch normalization and a dropout layer
	3,757,251 (tri-class)	
CNN	7138 (bi class)	Contains max_pooling2d layer, batch normalization, global averaging pooling2d and dense layers
	7203 (tri-class)	
Fine Tuned CNNs	11,954,851 (tri-class)	Contains max_pooling2d layer, batch normalization, dropout, and dense layers
	11,954,722 (bi class)	
Knowledge distillation (KDTL)	238,525,828 (bi class)	Combined the student and teacher model
	238,526,854 (tri-class)	
	238,823,562 (5 class)	

Table 4 Different DL models can be used within our KDTL framework. Accuracies (training/testing/validation) for different combinations of teacher and student networks are given along with the computational time (seconds). The models were tested on the Bonn University and ISRUC datasets for two and five-class classifications

Model	Dataset	Accuracy (training %)	Accuracy (validation %)	Accuracy (testing %)	Time taken (training)	Time taken (testing)
ConvNext + MobileNet	Bonn University (2 class)	100	100	98	377.09/265.00	31.17/0.54
ConvNext + CNN	Bonn University (2 class)	89.74	87.99	92	402.17/225.78	48.13/0.5912
ConvNext + fine-tuned CNN	Bonn University (2 class)	100	100	96	281.30/243.0	29.76/0.757
ConvNextTiny + MobileNet	Bonn University (2 class)	87.5	60	60	206.43/54.14	24.00/0.58
ConvNext + MobileNet	Bonn University (3 class)	98.5	98	98	389.54/247.40	39/0.84
ConvNext + CNN	Bonn University (3 class)	91	92.00	80	374.65/432.11	29.87/0.73
ConvNext + fine-tuned CNN	Bonn University (3 class)	100.0	93.99	98	434.83/168.88	46.83/0.741
ConvNextTiny + MobileNet	Bonn University (3 class)	82.99	40.00	46.4	84.60/50.85	0.945/0.464
Conv2Next + MobileNet	ISRUC Dataset (2-class)	96.07	95.12	95	361.05/154.33	3.43/0.802
ConvNext + CNN	ISRUC Dataset (2-class)	85.80	85	85.71	226.86/99.68	24.38/0.48
ConvNext + fine-tuned CNN	ISRUC Dataset (2-class)	100.0	97.5	100	385.52/283.315	46.92/0.73
ConvNextTiny + MobileNet	ISRUC Dataset (2-class)	78.24	85	85.71	209.65/64	47.19/0.71
Conv2Next + MobileNet	ISRUC Dataset (5-class)	99.27	77.68	77.49	33,197.14/ 22586.67	1048 s/11.76

and MobileNet combination obtained the highest training accuracy while in the binary classification, ConvNext and fine-tuned CNN performed well overall. Our testing also showed that the computational burden of the ConvNext and CNN models are lower than the Conv2Next with MobileNet. Our KDTL framework can further be tested with other DL model combinations and we next compare and contrast our approach from related models in the literature.

3.3 Comparison with other models

In [45, 46], authors have used SVM and have demonstrated high classification accuracy; however, CNNs can learn increasingly complex features at each layer, allowing them to capture both low-level and high-level features that are relevant to the task at hand. In [47], authors have demonstrated the use of EOG, EMG along EEG in classifying sleep stages using an SVM classifier. Additionally, SVMs only consider a fixed set of features that are handcrafted or engineered, which can limit their ability to capture the full complexity of the data. CNNs are also well suited for computer vision tasks such as image classification, where the input data can be highly variable. SVMs, on the other hand, are more sensitive to variations in input data and may require additional preprocessing or feature engineering to handle these variations. Additionally, SVM is also computationally intensive, on the other hand using knowledge distillation, we can use the lightweight model to give higher accuracy than the study in the paper. Finally, the knowledge distillation approach can be used to create a student model that is tailored to a specific task, whereas an SVM is a more general-purpose classifier that may not be as well suited to certain types of data or problems. Chandel et al. [48] have utilized KNN for the classification of sets; however, it should be noted that for binary classification complete dataset was not used to train the models; in comparison, complete data were used in our study for training our model in both binary and 3 classes. Also, MobileNet can learn a more generalized representation of the input data, which can improve its performance on unseen data compared to using KNN. This is because MobileNet can learn from the entire dataset, whereas KNN is based on the nearest neighbors of each test point, which may not be representative of the whole dataset. This can result in overfitting or underfitting, which can reduce the generalization performance of KNN. Rout et al. [49] introduced a new framework called multi-fused reduced CNN (MF-RDCNN) in which the first component is a set of convolutional layers that extract features from the input data. This is followed by max pooling and dropout layers to reduce overfitting. The second component is a set of fully connected layers that classify the features extracted by the convolutional layers. The third component is the multi-fuse architecture, which combines the outputs of multiple branches of the network. Each branch of the network has a different architecture, which includes different numbers of convolutional and fully connected layers. The outputs of the different branches are combined using a weighted averaging technique, where the weights are learned during training. The architecture requires a significant number of computational resources to train and evaluate. This may be a disadvantage in scenarios where computational resources are limited or where real-time processing is required.

In [50], authors have extracted a set of features from the EEG signals using an orthonormal discrete wavelet transform. These features are then fed into a fuzzy logic system, which uses a set of fuzzy rules to generate an output in the form of a Q-table. In the second stage, a Q-learning algorithm is used to optimize the Q-table generated in the first stage. A genetic algorithm is then used to select the best features and fine-tune the parameters of the fuzzy logic system. The final result is a classifier that can accurately identify seizure events in EEG signals. Knowledge distillation is a relatively simple approach that only requires the use of two models (a teacher and a student), whereas the genetic algorithm approach requires more complex algorithms and a larger computational cost. The use of fuzzy logic and genetic algorithms in the proposed method also makes it difficult to interpret how the algorithm arrives at its decisions, which could be a concern in some applications where interpretability is important. The study also did not analyze other classes and has only performed 5 class classifications. In [51], authors proposed a new method for epileptic seizure classification using variational mode decomposition (VMD) and an error-minimized random vector functional link (ERVL) neural network called EMRVFLN classifier. However, the EMRVFLN classifier may not provide easily interpretable results, as it relies on complex feature mapping and multiple hidden layers. In contrast, knowledge distillation using ConvNext as teacher and MobileNet as student is designed to extract knowledge from the teacher model and transfer it to the student model, resulting in a more interpretable and transparent model. Additionally, no regularization techniques are mentioned in the paper since EMRVFLN may suffer from overfitting when the number of neurons in the hidden layer is too large, which may require regularization techniques to prevent.

Wang et al. [44] proposed a multilayer graph attention network for sleep classification; in the preprocessing stage, the raw EEG signals are transformed into a set of frequency bands using a wavelet transform. These frequency bands are then used to construct a graph structure where each node corresponds to a specific EEG channel, and the edges represent the correlations between the channels. The model is trained on the graph structure, and it uses graph attention

Table 5 Comparison of state-of-the-art models from the literature

Dataset	Authors	Year	Method	Results
Bonn University dataset	[45]	2021	SVM, KNN	99.6%
Bonn University dataset	[48]	2019	KNN	99.45%
Bonn University dataset	[49]	2022	Multi-fuse reduced deep convolutional neural network	99.82%
Bonn University dataset	[46]	2021	Multiclass support vector machine	99.59%
Bonn University dataset	[50]	2021	Reinforcement learning techniques of FQL and GA FQL	91%, 94.6%
Bonn University dataset	[51]	2020	Error-minimized random vector functional link network using least-square support vector machine and extreme learning machine	100% (2-class), 99.74% (3-class)
Bonn University dataset	[54]	2021	Separate Tunable Q Wavelet Transform level decomposition for filtering, Multiclass SVM	99.59%, 100%
Bonn University dataset	[55]	2020	CNN	98.65%
Bonn University	Ours	2023	KD with ConvNext (teacher) Mobilnet (student)	100% (2 class) 98% (3 classes)
ISRUC-Sleep	[44]	2022	Multilayer graph attention network (MGANet)	82.5% (5 class)
ISRUC-Sleep	[52]	2022	Graph neural network (GNN)	87.4% (5 class)
ISRUC-Sleep	[56]	2022	SVM DeepSleepNet	57.9% (5 class)
ISRUC-Sleep	[47]	2022	Random Forest	80.8% (5 class)
ISRUC-Sleep	[53]	2021	SVM	90% (2 class)
ISRUC-Sleep	Ours	2023	KD with ConvNext (teacher) Mobilnet (student)	97.28% (2 class) 75% (5 class)

layers to learn the nonlinear relationships between the EEG signals across multiple channels. The attention mechanism allows the model to focus on the most informative EEG channels for each sleep stage classification task. The total number of parameters is 1.5×10^5; the number of epochs used to train the model was 80; however, in our case, using early stopping, we required 9 epochs to train the model. Li et al. [52] proposed a deep learning model for sleep classification which uses spatiotemporal graph convolutional networks (ST-GCN). The architecture consists of nine ST-GCN modules, each followed by an attention (ATT) block. Each ST-GCN module is made up of a graph convolutional network (GCN) block followed by a temporal convolutional network (TCN) block. The GCN block is used to capture the spatial relationships between different EEG channels, while the TCN block is used to capture the temporal dynamics of the signal. The binary classification in our model gives a high accuracy and is well suited to classify between sleep disorder and healthy patients additionally is less complex than ST-GCN, it also utilizes 120 epochs to train the model. Jayaraj et al. [53] proposed a method for the classification of sleep apnea based on the sub-band decomposition of electroencephalography (EEG) signals. For classification, SVM and random forest were used. However, random forest may give more importance to certain features that are highly correlated with the target variable, while ignoring other important features that may not be highly correlated. This can be a limitation when

dealing with high-dimensional data like EEG where feature selection is important. Moreover, in our study, we have used spectrograms, which can be thought of as two-dimensional images. This allows our model to leverage its ability to learn hierarchical representations of the data, which can capture both local and global features. On the other hand, random forest is a decision tree-based algorithm that relies on handcrafted features for classification. While random forest can perform well on small datasets, it may struggle with high-dimensional data such as EEG signals and may not be able to learn complex features from raw data. We have also demonstrated a reduction in the number of parameters along with time while maintaining accuracy. For our proposed combination, i.e., ConvNext and MobileNet, we have seen a reduction in time taken from 30 to 0.5 s.

Table 5 shows the comparison of state-of-the-are methods versus our KDTL framework on both the Bonn University and ISRUC-Sleep datasets. We outperform other DL models on both the datasets, and our best model with ConvNext/MobileNet performed well overall. From the above results, it can be seen that the results of the paper can beat many of the previous attempts made. While our binary classification model achieved a 100% validation accuracy on the Bonn dataset, it is important to note that the limited size of the data and the complexity of the models used may have contributed to this high accuracy. In addition to using transfer learning-based models that were pretrained on a larger dataset and fine-tuned on our dataset, we also used various normalization techniques such as batch normalization, dropout layers, and L1–L2 regularization to reduce overfitting. However, upon using L1–L2 regularization on the teacher model with values of 0.1 and 0.01 we noticed that the accuracy dropped to 87% for the teacher model, this could be because EEG signals can have highly variable amplitude and frequency spectra. Despite these limitations, our framework has produced promising results on the datasets, and we plan to further test it on larger datasets with more patients in the future. We have also compared our methodology to other combinations and found that our approach outperforms several other techniques. Overall, while our results are encouraging, we recognize the need for further validation and testing on more diverse datasets. Our goal is to continue refining our approach and improving its accuracy and reliability for the diagnosis of epilepsy using EEG signals.

4 Conclusions

In this work, we have discussed the classification of various mental disorders using time–frequency representation of the EEG signals and deep learning (DL) models. The main motive of this paper was to find a solution which can be applied to real-life clinical applications; hence, we utilized a lightweight model in our investigations. The paper has demonstrated the results based on two different datasets. The main contribution mentioned in this paper is the novel methodology of using knowledge distillation which has demonstrated a reduction of a number of parameters along with the reduction in overall training time and obtained good accuracies across two different datasets, namely the Bonn University, and ISRUC. We also showed that our knowledge distillation transfer learning (KDTL) framework can be tested with various combinations of the DL models and we have provided experimental comparisons. Compared to previous models in the literature our framework obtained better results in binary and multiclass classifications. Further, our proposed approach can be amenable to potential clinical applications; however, this requires further validation to check for model drift as new EEG data are acquired. Though we have made considerable progress in our work some of the challenges to be addressed are: (1) testing the proposed methodology on a greater number of datasets to check its validity and generalizability; even though we have tested on two different datasets and on multiple cases, there are further scenarios wherein our KDTL framework can be further tested; so in the future, more datasets can be tested using our methodology. (2) Testing on more advanced models: For our research, we have used the MobileNet model as our student model and ConvNext as the teacher model and obtained highly satisfactory results, in the future more advanced models can be tested and investigated to reduce the number of parameters and training time further.

Funding None.

Data availability The datasets generated during and/or analyzed during the current study are available in the Bonn repository, https://www.ukbonn.de/epileptologie/arbeitsgruppen/ag-lehnertz-neurophysik/downloads and ISRUC repository, https://sleeptight.isr.uc.pt.

Declarations

Conflict of interest The authors declare no conflict of interests.

References:

1. Ullah H et al (2019) Internal emotion classification using EEG signal with sparse discriminative ensemble. IEEE Access 7:40144–40153
2. Manasa G et al (2024) EEG signal-based classification of mental tasks using a one-dimensional ConvResT model. Neural Comput Appl 36:1–20

3. Shanmugam S, Dharmar S (2023) A CNN-LSTM hybrid network for automatic seizure detection in EEG signals. Neural Comput Appl 35(28):20605–20617
4. Gao Z et al (2024) FAformer: parallel Fourier-attention architectures benefits EEG-based affective computing with enhanced spatial information. Neural Comput Appl 36(8):3903–3919
5. Badr Y et al (2024) A review on evaluating mental stress by deep learning using EEG signals. Neural Comput Appl 36:1–26
6. Bhattacharyya A, Pachori RB (2017) A multivariate approach for patient-specific EEG seizure detection using empirical wavelet transform. IEEE Trans Biomed Eng 64(9):2003–2015
7. Das AB, Bhuiyan MIH (2016) Discrimination and classification of focal and non-focal EEG signals using entropy-based features in the EMD-DWT domain. Biomed Signal Process Control 29:11–21
8. George FP et al (2019) Recognition of emotional states using EEG signals based on time-frequency analysis and SVM classifier. Int J Electr Comput Eng 9(2):2088–8708
9. Zheng J et al (2021) Time-frequency analysis of scalp EEG with Hilbert-Huang transform and deep learning. IEEE J Biomed Health Inform 26(4):1549–1559
10. Tabar YR, Ugur H (2016) A novel deep learning approach for classification of EEG motor imagery signals. J Neural Eng 14(1):016003
11. Madhavan S, Tripathy RK, Pachori RB (2019) Time-frequency domain deep convolutional neural network for the classification of focal and non-focal EEG signals. IEEE Sens J 20(6):3078–3086
12. Tawhid MNA, Siuly S, Hua W (2020) Diagnosis of autism spectrum disorder from EEG using a time–frequency spectrogram image-based approach. Electron Lett 56(25):1372–1375
13. Zhang C et al (2022) multichannel multidomain-based knowledge distillation algorithm for sleep staging with single-channel EEG. IEEE Trans Syst II Exp Br 69(11):4608–4612
14. Khan NA et al (2021) A novel knowledge distillation-based feature selection for the classification of ADHD. Biomolecules 11(8):1093
15. Gou J et al (2021) Knowledge distillation: a survey. Int J Comput Vision 129(6):1789–1819
16. Zhang G, Etemad A (2021) Distilling EEG representations via capsules for affective computing. Pattern Recognit Lett 171:99–105
17. Ieracitano C et al (2020) A novel multi-modal machine learning based approach for automatic classification of EEG recordings in dementia. Neural Netw 123:176–190
18. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
19. He, K., et al. (2016) "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition
20. Liu Z, et al. (2022) A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition
21. Tao Y, Chang F, Huang Y, Ma L, Xie L, Su H (2022) Cotton disease detection based on ConvNeXt and attention mechanisms. IEEE J Radio Freq Identif. https://doi.org/10.1109/JRFID.2022.3206841
22. Fan S et al (2023) LACN: A lightweight attention-guided ConvNeXt network for low-light image enhancement. Eng Appl Artif Intell 117:105632
23. Gu J et al (2018) Recent advances in convolutional neural networks. Pattern Recognit 77:354–377
24. He K, et al. (2017) Mask R-CNN. Proceedings of the IEEE international conference on computer vision
25. Ho TKK, Gwak J (2020) Utilizing knowledge distillation in deep learning for classification of chest X-ray abnormalities. IEEE Access 8:160749–160761
26. Ibrahim S, Djemal R, Alsuwailem A (2018) Electroencephalography (EEG) signal processing for epilepsy and autism spectrum disorder diagnosis. Biocybern Biomed Eng 38(1):16–26
27. Jia Mi et al (2022) KDE-GAN: A multimodal medical image-fusion model based on knowledge distillation and explainable AI modules. Comput Biol Med 151:106273
28. Bi C et al (2020) MobileNet based apple leaf diseases identification. Mob Netw Appl 27:1–9
29. Wang W et al (2020) A novel image classification approach via dense-MobileNet models. Mob Inform Syst. 2020:1
30. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) OverFeat: integrated recognition, localization and detection using convolutional networks. arXiv:1312,6229.
31. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
32. Howard AG, et al. (2017) MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861
33. Yang J et al (2022) MobileNet and knowledge distillation-based automatic scenario recognition method in vehicle-to-vehicle systems. IEEE Trans Veh Technol 71(10):11006–11016
34. Chiu YC, et al. (2020) Mobilenet-SSDv2: An improved object detection model for embedded systems. International conference on system science and engineering (ICSSE). IEEE
35. Sun Y, Zhang J, Chaoyue H (2021) A flower recognition system based on MobileNet for smart agriculture. IEEE 3rd international conference on frontiers technology of information and computer (ICFTIC). IEEE
36. Yin H, et al. (2020) Dreaming to distill: Data-free knowledge transfer via deepinversion. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
37. Malinin A, Mlodozeniec B, Gales M (2019) Ensemble distribution distillation. arXiv preprint arXiv:1905.00076
38. Cho JH, Hariharan B (2019) On the efficacy of knowledge distillation. Proceedings of the IEEE/CVF international conference on computer vision
39. Heo B, et al. (2019) A comprehensive overhaul of feature distillation. Proceedings of the IEEE/CVF international conference on computer vision
40. Gao M et al. (2020) Residual knowledge distillation. arXiv preprint arXiv:2002.09168
41. Kulkarni A et al. (2019) Data efficient stagewise knowledge distillation. arXiv preprint arXiv:1911.06786
42. Shen Z, He Z, Xue X (2019) Meal: Multi-model ensemble via adversarial learning. Proc AAAI Conf Artif Intell. 33(01):4886
43. Aguilar G et al (2020) Knowledge distillation from internal representations. Proc AAAI Conf Artif Intell. 34(05):7350
44. Wang Q et al (2022) Multi-layer graph attention network for sleep stage classification based on EEG. Sensors 22(23):9272
45. Yazid M et al (2021) Simple detection of epilepsy from EEG signal using local binary pattern transition histogram. IEEE Access 9:150252–150267
46. Agrawal R, Bajaj P (2021) Comparative classification techniques for identification of brain states using TQWT decomposition. J Intell Fuzzy Syst 41(5):5287–5297
47. Li Y, et al. (2022). Automatic sleep stage classification based on two-channel EOG and one-channel EMG. Physiological Measurement
48. Chandel G et al (2019) Detection of seizure event and its onset/offset using orthonormal triadic wavelet based features. IRBM 40(2):103–112
49. Rout SK et al (2022) An efficient epileptic seizure classification system using empirical wavelet transform and multi-fuse reduced

deep convolutional neural network with digital implementation. Biomed Signal Process Control 72:103281
50. Kukker A, Sharma R (2021) A genetic algorithm assisted fuzzy Q-learning epileptic seizure classifier. Comput Electr Eng 92:107154
51. Rout SK, Biswal PK (2020) An efficient error-minimized random vector functional link network for epileptic seizure classification using VMD. Biomed Signal Process Control 57:101787
52. Li M, Chen H, Cheng Z (2022) An attention-guided spatiotemporal graph convolutional network for sleep stage classification. Life 12(5):622
53. Jayaraj R, Mohan J (2021) Classification of sleep apnea based on sub-band decomposition of EEG signals. Diagnostics 11(9):1571
54. Kaushik G et al (2022) EEG signal based seizure detection focused on Hjorth parameters from tunable-Q wavelet sub-bands. Biomed Signal Process Control 76:103645
55. Woodbright M, Verma B, Haidar A (2021) Autonomous deep feature extraction based method for epileptic EEG brain seizure classification. Neurocomputing 444:30–37
56. Ye J et al (2021) CoSleep: A multi-view representation learning framework for self-supervised learning of sleep stage classification. IEEE Signal Process Lett 29:189–193
57. Bhattacharyya A et al (2018) A novel approach for automated detection of focal EEG signals using empirical wavelet transform. Neural Comput Appl 29:47–57